The Double Beta Model for Proportion Data Composed of Heterogeneous Binary Responses *

Kentaro Fukumoto †

July 1, 2005

Abstract

When political scientists observe proportion as average of binary responses, they are sometimes interested in how heterogeneously probabilities of these unobserved individual binary responses are distributed. The double beta model I propose here enables them to infer this by assuming beta distribution both at individual level and collective level. Monte Carlo simulation shows how accurate and robust this estimator is and compares it with the log odds normal models. Also, as an example of empirical application, I reanalyze a study of legislative gridlock.

*I appreciate Sarah Binder for providing her data. This is work in progress. Comments are really welcome.

†Visiting Scholar, Reischauer Institute of Japanese Studies, Harvard University, Cambridge, MA, and Associate Professor, Department of Political Science, Gakushuin University, Tokyo, E-mail: Kentaro.Fukumoto@gakushuin.ac.jp, URL: http://www-cc.gakushuin.ac.jp/~e982440/index_e.htm.

1 Introduction

Political scientists very often analyze proportion as average of binary responses. Examples are approval rate of the executive chief and percentage of party identifiers at survey, share of votes in elections and seats in the legislature, unemployment rate, legislative productivity, viewing rate of campaign ads programs on TV and so on.

Most researchers implicitly assume that a proportion variable follows the normal distribution and employs identity link for the expectation. This model, sometimes inaccurately called OLS, is not appropriate because it does not take into consideration the fact that proportion varies only between 0 and 1. Besides, it follows that distribution of a proportion is heteroskedastic.

To address this problem, some statisticians (including political methodologist) propose the beta regression and the log odds normal (Paolino, 2001). Both use logit link for the expected value of proportion so that linear predictor has no boundary. Beta regression assumes that proportion is beta distributed and the log odds normal model supposes that log odds of proportion is a normal random variable.

These methods are promising, though they do not take seriously binary responses of which proportion is composed. But probabilities that binary responses are positive are sometimes of real interest for political science. For example, apparently, voters do not have the same probability to support the government. Republicans should have higher probability to approve a Republican president than Democrats. That is, binary response is heterogeneous. If Republicans always approve a Republican president and Democrats always disapprove him, approval rate should be constant as long as the party identification rate remains the same. As Independent voters, who have the probability of 50% , increase, namely, as voters become homogeneous, approval rate will get volatile. This simple example briefly shows that, even if expected value of individual probabilities is the same, different

distribution of individual probabilities affects distribution of collective proportion. Other examples of individual probability are that of incumbent victory (safe seat or competitive seat) and polarization of legislators in ideologies.[1]

This paper intends to infer how unobserved individual probabilities are distributed from observed collective proportion data. The double beta model I propose here assumes beta distribution both at individual level and collective level. One can estimate distribution of individual probabilities once the number of binary responses for each proportion value is known.

The remainder of this paper is composed as follows. The next section explains the data generation process of the double beta model and its estimator. In the third section, Monte Carlo simulation shows how accurate and robust this estimator is for various values of parameters and compares the double beta model with the log odds normal models. Then, as an example of empirical application, I reanalyze a result in Binder (2003). Finally, I conclude.

2 Model

2.1 Data Generation Process

A random dependent variable $0 < P < 1$ is a proportion of individuals whose latent (unobserved) binary responses, $Z_i = \{0, 1\}(i = 1, \ldots, n)$, are positive ($z_i = 1$) among n individuals;

$$P = \frac{1}{n} \sum_{i=1}^{n} Z_i$$

[1] Palmquist (1997) tries to consider individual level heterogeneity for count data. Extended beta-binomial distribution he adopts, however, can not model it completely and tell how distributed individual probabilities are.

Consider an example of approval rate. When, among sample size $n = 1000$ interviewees, 600 people approve the government ($z_i = 1$) and 400 do not ($z_i = 0$), approval rate p is 600/1000=0.6.

The individual binary response Z_i independently follows (not identical) Bernoulli distribution whose expectation is $0 < \pi_i < 1$;

$$Z_i \sim \mathcal{BR}(z_i|\pi_i) = \pi_i^{z_i}(1-\pi_i)^{1-z_i}$$

For instance, if a respondent support the government strongly, π_i will be high, say, 0.9. Unless something bad happens, this person will answer that he or she approves the government. By contrast, when one dislikes the government, π_i should be low like 0.1 and that individual rarely expresses approval if asked. Aside from them, those who wonder if the government is good or bad take the value of π_i near 0.5 and flip-flop at poll. Thus, individual probability π_i means his or her tendency to approve the government.

I construct π_i's so that they are ideally stratified sample draws from a beta random variable $0 < \pi < 1$ whose mean and dispersion parameter are $0 < \bar{\pi} < 1$ and $\gamma > 0$, respectively. Or, the cumulative distribution function of π_i's is approximated by that of beta random variable. Rearrange π_i's so that $\pi_i \geq \pi_j$ for $i > j$. π_i is expectation of sample draws from the ith part of n equally divided quantile segments of beta distributed π;

$$\begin{aligned}
\pi &\sim \mathcal{BT}(\pi_0|\bar{\pi},\gamma) \\
&= \frac{\Gamma(\gamma^{-1})}{\Gamma(\bar{\pi}\gamma^{-1})\Gamma([1-\bar{\pi}]\gamma^{-1})}\pi_0^{(\bar{\pi}\gamma^{-1})-1}(1-\pi_0)^{([1-\bar{\pi}]\gamma^{-1})-1} \\
\pi_i &= E\left[\pi_0 \middle| Q\left(\pi \middle| \frac{i-1}{n}\right) < \pi_0 < Q\left(\pi \middle| \frac{i}{n}\right)\right] \\
&= n\int_{\frac{i-1}{n}}^{\frac{i}{n}} \mathcal{IBT}^{-1}(q)dq
\end{aligned}$$

where $Q(X|q)$ is the q quantile value of random variable X and $\mathcal{IBT}(\pi_0)(0 \leq \pi_0 \leq 1)$ is the incomplete beta function (the cumulative distribution function for the beta random variable);

$$\mathcal{IBT}(\pi_0) = \int_0^{\pi_0} \mathcal{BT}(\pi) d\pi$$

Large γ leads to large variance of π and most of π_i's approach 0 or 1. Figure 1 (1) visualizes an example where $\bar{\pi} = 0.5$ and $\gamma = 2$ as a heterogeneous case. If P is an approval rate, citizens are thought to been polarized between sympathizers and antagonists toward the government. If γ is small, say, 0, π has small variance and many π_i's shrink to their average $\bar{\pi}$ (a homogeneous case in the figure). Then, voters are almost eqaully ambivalent.

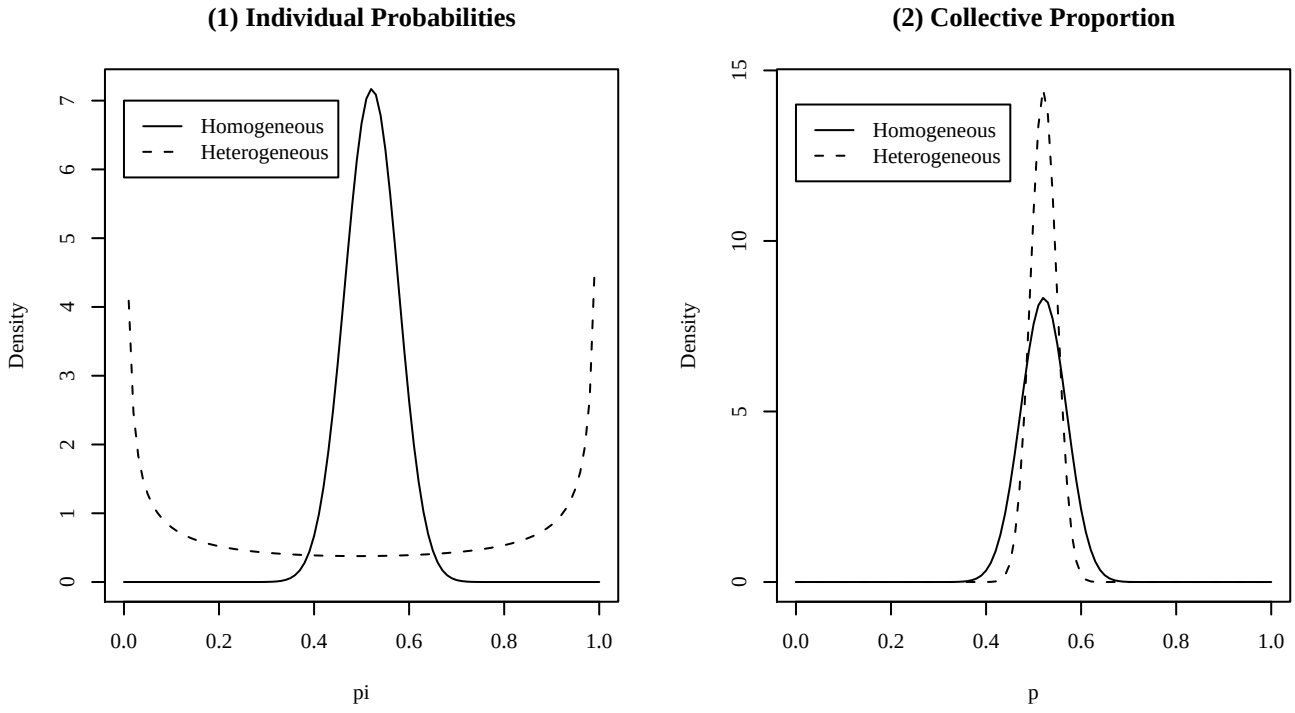

Figure 1: Individual Probabilities and Collective Proportion in the Cases of Homogeneous and Heterogeneous Individuals

From above, asymptotic expectation and variance of collective proportion P are;

$$\lim_{n\to\infty} E(P) \to \bar{\pi}$$
$$\lim_{n\to\infty} V(P) \to \frac{\bar{\pi}(1-\bar{\pi})}{n(1+\gamma)}$$

(For derivation, see Appendix (3).) Hence, expectation of collective proportion is approximated by mean of individual probability. Also, in the case of approval rate, where most people's mind are decided (large γ, thus, most of π_i's are near 0 or 1), approval rate P is centered on its expected value $\bar{\pi}$ and less volatile (small variance) (see Figure 1 (2)). On the other hand, when there are more undecided voters (small γ, thus, more π_i's are near their mean $\bar{\pi}$), variance of approval rate P becomes wider. Though this twisted nature of variances at the two levels may be misleading, this relationship sounds reasonable.

If I assumed all π_i's are the independent and identical random variable as π and sample π_i's randomly from π, large precision of decided mind voters' collective proportion described above (small $V(P)$) would be exactly canceled out by large sampling variance due to individual heterogeneity among people (large $V(\pi)$, or vice versa), which we can not detect from data. In order to avoid this inconvenience, instead of assuming that population densities of π_i's follow beta distribution, I approximate sample cumulative probability mass function of π_i's by cumulative beta function. Thus, strictly speaking, this model does not tell us how distributed individual probabilities π_i's in the population are. I believe, however, that population cumulative distribution function is not so different from sample one anyway and the ability of this model to examine heterogeneity among citizens well overwhelms, if any, demerits.

2.2 Estimator

My quantities of interest are expectation and dispersion parameters of the individual probabilities π_i's' distribution, $\bar{\pi}$ and γ, rather than any parameter of the collective proportion P's distribution. Thus, I reparameterize $\bar{\pi}$ and γ by using covariates x and w, respectively;

$$\bar{\pi} = \frac{1}{1+\exp(x\beta)} \qquad (0 < \bar{\pi} < 1)$$
$$\gamma = \exp(w\alpha) > 0$$

We can not, however, observe π_i's, much less z_i's, but only p. Hence, we should estimate β and α by using p.

Here I approximate the distribution of P by the beta probability density function. For sufficiently large n, we can regard;

$$P \sim \mathcal{BT}\left(p\middle|\bar{\pi}, \frac{1}{n(1+\gamma)-1}\right)$$

(For derivation, see Appendix (3).) When there are T observations of p_t's, likelihood $\mathscr{L}$ is;

$$\begin{aligned}\mathscr{L}(\mathbf{p}) &= \prod_{t=1}^{T} \mathscr{L}_t(p_t) \\ &\propto \prod_{t=1}^{T} \mathcal{BT}\left(p_t\middle|\frac{1}{1+\exp(x_t\beta)}, \frac{1}{n_t[1+\exp(w_t\alpha)]-1}\right)\end{aligned}$$

Note that the number of observations T (e.g. 60 observations in 5 years monthly survey data) is different from the number of samples in the tth observation, n_t (e.g. every survey may have 1000 respondents). Then, one can estimate parameters β and α by maximum likelihood method.

3 Monte Carlo Simulation

3.1 Various Values of Dispersion Parameter γ

First, I check accuracy and robustness of the double beta estimator. It does not work well when the true value of individual dispersion parameter γ is too small or large, namely, π_i takes almost only one value (e.g. completely homogeneous respondents) or the two values of 0 and 1 (e.g. mind decided supporters and dissidents). I show this by Monte Carlo simulation.

Covariates x (for mean) and w (for dispersion) are constant term and a random draw from standard normal distribution. Parameters for mean are $\beta = (\beta_0, \beta_x)' = (1, -1)'$ and those for dispersion are $\alpha = (\alpha_0, \alpha_w)' = (\alpha_0, 0)'$, where I change constant for dispersion α_0 from -4 to 10 by 1. Sample size n_t's are random draws from Poisson variable with mean equal to 1,000. $T = 50$ observations of y are generated for every value of α_0.

You find summary of estimates of α_0 and α_w in Table 1.[2] Though estimates of α_0 is biased except for $\alpha_0 = 5$, their mediana are close enough to the true values for $1 \leq \alpha_0 \leq 6$. The upper bounds of 95% confidence intervals (the 97.5% quantile) are not so far from the true values for $0 \leq \alpha_0 \leq 6$. Their lower bounds (the 2.5% quantile) are satisfactory only for $4 \leq \alpha_0 \leq 6$. When it comes to estimates of α_w, their medians are always near their true values. They are, however, biased for $\alpha_0 \leq 2$ and their 95% confidence sets are too wide for $\alpha_0 \leq 3$. Therefore, when the true value of γ is less than e^1 or larger than e^6, estimates of α are not reliable. I suspect that, in these situations, it is numerically impossible to identify γ.

When γ is very small (π_i's are almost homogeneous), we can use the (beta) binomial model. If γ is very large (π_i's are almost fixed at either 0 or 1), proportion p has underdispersion and the extended beta binomial model captures the data generation process to some

[2] The statistics software `R` is used for estimation.

True	Dispersion's Constant, α_0				Coefficient of x, $\alpha_w = 0$			
α_0	Mean	Median	95% CI		Mean	Median	95% CI	
-3	−14.75	−8.32	−58.89	−0.76	−1.67	−0.31	−29.88	22.11
-2	−15.58	−7.99	−60.09	−0.40	−0.75	−0.26	−39.64	27.84
-1	−12.64	−5.96	−64.12	−0.03	−1.77	−0.09	−48.70	22.19
0	−16.17	−9.25	−91.26	0.59	−3.24	−0.33	−51.94	17.62
1	−14.65	0.63	−89.94	1.58	−2.90	−0.09	−45.83	14.59
2	−9.16	1.95	−69.13	2.47	−2.16	−0.04	−32.71	15.50
3	−0.14	3.03	−26.95	3.39	0.33	0.03	−0.88	5.78
4	3.32	4.02	3.59	4.41	−0.32	0.01	−0.49	0.45
5	4.99	4.98	4.62	5.43	−0.08	−0.08	−0.58	0.39
6	5.73	5.73	5.36	6.20	−0.10	−0.11	−0.47	0.26
7	5.83	5.97	5.68	6.39	−0.02	−0.12	−0.42	0.17
8	6.06	6.06	5.74	6.40	−0.13	−0.12	−0.48	0.15
9	6.07	6.05	5.82	6.43	−0.12	−0.11	−0.51	0.17
10	6.05	6.03	5.76	6.39	−0.11	−0.13	−0.52	0.22

Table 1: Monte Carlo Simulation of Estimates of α by Different True Values of Dispersion's Constant α_0

degree (Palmquist, 1997). I will argue these alternative models in another paper.

I save the space by not showing results of β_0 and β_x because their biases are less than 0.002 and their standard deviances are less than 0.02. Thus, estimates of β are satisfactory in almost any situations.

3.2 Comparison with the Log Odds Normal Models

Next, what do we earn by using the double beta model? Obviously, we can know not only the distribution of collective proportion but also that of individual probabilities. The second Monte Carlo simulation compares estimates of the present model with one of the promising rival models, the log odds normal model.

The log odds normal model is the easiest way to take into consideration definite range and heteroskedastic nature of proportion data. It assumes that log odds of proportion p

follows normal distribution.

$$\log\left(\frac{p}{1-p}\right) \sim \mathcal{N}(p|x\beta, \gamma)$$

You can perform this on ready made software by taking log odds of proportion and using it as dependent variable of the normal regression. This is a homoskedastic model in the sense that the normal distribution, not proportion, has constant variance. If we model this variance by covariates ($\gamma = \exp(w\alpha)$), we have a heteroskedastic model.

The simulation setting is the same as the previous one except that parameters for dispersion are $\alpha = (\alpha_0, \alpha_w) = (2, 1)$ and sample size n_t's are constantly 1,000. The results are summarized in Table 2. As for β, all of the three estimators are almost equally unbiased and efficient.[3] Estimates of α by the double beta are volatile, though its median is close to the true value. Since the heteroskedastic log odds normal model parameterizes dispersion in different way from the double beta, we can not compare α_0. Also, α_w is expected to be negative and it turns out to be so, significantly different from zero. But here comes how the two models are different. The log odds normal models can not show how distributed individual probabilities π_i's are because they do not utilize information of sample size n. Even if you include sample size n as covariates of dispersion γ, the estimator can not identify constant term for dispersion (α_0) and the coefficient of n (α_n), because both have multicollinearity problem. Hence, if researchers are really interested in not only the distribution of collective proportion but also that of individual probabilities, the double beta model is recommended.

Though I also performed estimation by the beta regression and the variance constrained normal model with logit transformation,[4] these lead to similar conclusion. Thus, I do not

[3] Paolino (2001) argues that first difference of log odds normal model is less efficient than that of beta regression.

[4] Suppose Z_i follows identically and independently Bernoulli distribution $\mathcal{BR}(z_i|\bar{\pi})$. Then, from the central limit theorem and Appendix (2), asymptotically;

$$\begin{aligned} p &\sim \mathcal{N}(p|E(P) = E(Z_i), V(P) = V(Z_i)/n) \\ &= \mathcal{N}\left(p\middle|\bar{\pi} = \frac{1}{1+\exp(-x\beta)}, \frac{\bar{\pi}(1-\bar{\pi})}{n}\right) \end{aligned}$$

	Mean	Median	S.D.	RMSE
Parameters for Mean				
Constant ($\beta_0 = 1$)				
Double Beta	0.99970	0.99952	0.00409	0.00408
Heteroskedastic Log Odds Normal	0.99996	1.00009	0.00380	0.00378
Homoskedastic Log Odds Normal	1.00077	1.00056	0.00528	0.00531
Coefficient for x ($\beta_x = -1$)				
Double Beta	−0.99872	−0.99877	0.00462	0.00477
Heteroskedastic Log Odds Normal	−0.99949	−0.99868	0.00556	0.00556
Homoskedastic Log Odds Normal	−1.00003	−0.99962	0.00617	0.00614
Parameters for Dispersion				
Constant ($\alpha_0 = 2$)				
Double Beta	−7.71445	1.94040	23.65744	
Heteroskedastic Log Odds Normal	−3.57161	−3.58051	0.12802	
Homoskedastic Log Odds Normal	0.00111	0.00108	0.00037	
Coefficient for w (α_w) = 1				
Double Beta	2.54095	1.11309	8.19940	
Heteroskedastic Log Odds Normal	−0.46349	−0.46744	0.15441	

Table 2: Monte Carlo Simulation of Estimates by Different Estimators

report those results.

4 Empirical Application: Heterogeneous Gridlock Tendency of Legislative Agendas

In order to show how useful the double beta model is for empirical application, I reanalyze Sarah Binder's *Stalemate: Causes and Consequences of Legislative Gridlock* (2003), especially, "the central analysis of the book," Table 4-2 (p. 68, 155). The question this book answers is what affects legislative gridlock rates in American politics. Her greatest contri-

Considering p is proportion, p should have the truncated normal distribution;

$$p \sim \mathcal{TN}_0^1\left(p \middle| \frac{1}{1+\exp(-x\beta)}, \frac{\bar{\pi}(1-\bar{\pi})}{n}\right)$$

This is what I call the variance constrained normal model.

bution is to conceptualize gridlock as the ratio of failed important agendas to all important ones, not negative number of successful agendas as previous works use it. This is dependent variable.[5] Independent variables are (see ch. 2 and Appendix E of her book);

- Divided government (dummy, positive coefficient is expected)
- Partisan moderation (number of moderates divided by distance between party medians, negative)
- Bicameral difference (percentage difference in voting yea on conference reports voted on by both chambers, positive)
- Time out of majority (average number of Congresses out of the majority when a new majority gains control of both chambers of Congress. The longer it waits, the more mandate it passes. positive)
- Budgetary situation (surplus or deficit divided by total federal outlays, negative)
- Public mood (lagged, Stimson policy mood indicator (public support for governmental action), negative)

Unit of analysis is Congress (biannual). The original data is published in the book (Appendix E). I succeed in replicating her own result by using weighted least square estimation for grouped data with logit function (Table 3, the left column).[6]

Then, I fit my double beta model. The same covariates as the author's are used for mean parameters. Besides, I expect that partisan moderation decreases individual dispersion parameter γ. Individual probability π_i means how difficult it is for that agenda to pass Congress. When lawmakers are polarized in their ideology (low partisan moderation), there

[5]Though she examines five indicies dependent on importance cutting points, the present paper considers only "gridlock 1" (to include important agendas most comprehensively).

[6]The author writes her data on the website, though I fail to find it. The author kindly gives it me upon my request. I appreciate her. I employ `glogit` command on STATA as she does.

	glogit		Double Beta	
Divided	0.306	*	0.252	**
government	(0.122)		(0.036)	
Moderation	0.006		−0.009	**
	(0.004)		(0.002)	
Bicameral	5.867	*	4.854	**
difference	(2.750)		(0.878)	
Time out of	0.010		−0.003	
majority	(0.024)		(0.004)	
Budgetary	0.005		0.005	*
situation	(0.007)		(0.002)	
Public mood	−0.007		0.002	**
(lagged)	(0.012)		(0.000)	
Constant	0.100		−0.203	
(mean)	(0.790)		(0.139)	
Constant			10.000	**
(dispersion)			(0.475)	
moderation			−0.462	**
			(0.065)	
N	24		24	

Table 3: Replication and Reanalysis of Binder's *Stalemate* (2003), Table 4-2, "Estimating the Frequency of Gridlock, 1953-2000" (Gridlock 1). Entries are estimates and standard errors are in parenthesis.

are little room for compromise or negotiation, you can easily predict that each agenda will be dead or alive and gridlock proportion p is less volatile (π_i's come near 0 or 1, γ as well as variance of π become large, and variance of p decreases). On the other hand, as more Congressmen stand in the center (high partisan moderation), fates of more agendas get uncertain and depend on maneuvers of these moderates and gridlock proportion p fluctuates more (π_i's converge around its mean $\bar{\pi}$, γ as well as variance of π get small, and variance of p is multiplied).

The result is shown in the right column of Table 3.[7] Compared with the original result, standard errors shrink very much. As a result, more coefficients turn out to be significantly

[7]For optimization on `R`, `genoud` command is used, because `optim` command returns negative curviture for the constant coefficient in α. In simulation above, `optim` is used.

different from zero and sometimes change their signs. Also, partisan moderation makes individual agenda's failure probabilities less volatile as expected. This contrast shows well that my double beta model enables researchers to get not only more precise estimates but also additional information about unobserved heterogeneity of individual response probabilities.

Actually, Figure 1 we saw before is drawn using this analysis result. It shows the density of individual agenda's gridlock probabilities and all agenda's gridlock proportion where the expected value of gridlock rates p, $\bar{\pi}$, is set at the mean of observed values (0.52). The homogeneous case and the heterogeneous one are where the variable moderation takes its mean (31.2) and the first quantile values (20.1), respectively (in these cases, γ's are about 0.01 and 2.02 and, according to the previous section, the estimation for this range γ's is reliable). When less legislators are moderate, most of agendas reduce to done deals and individual gridlock probabilities are either 0 (certainly pass) or 1 (certainly die) (Panel (1)). As a result, the probable range of collective gridlock proportion shrinks (Panel (2)). By contrast, when the proportion of moderates in Congress increases to the average level, many agendas have 50-50 individual gridlock probability (because more lawmakers wonder) and collective gridlock proportion spreads out. In the context of contemporary American politics, this analysis implies that polarized Congress deprives uncertainty, i.e. "the art of possibility", of important agendas. This way, the double beta model reveals latent micro level feature of the political phenomenon.

5 Conclusion

When is it worth employing the double beta model instead of other models? When researchers are interested in how individual binary response probabilities π_i's are distributed (homogeneous or heterogeneous) but they observe collective proportions only, not individual binary responses.

There are, however, some demerits. You have no canned software program, though it is not so hard to write a program. The number of respondents for every proportion n_t must be available, while it is often so. When respondents are too homogeneous or too heterogeneous, the present estimator does not behave well. In another paper, I will argue that, in these cases, you can use the (extended beta) binomial distribution.

It is easy and worth to overcome these demerits, while interest in heterogeneity of individual probabilities will grow in political science. Thus, I believe that the double beta model will become one of tool kits for scholars in the field.

Appendix: Mean, Variance and the Approximate Beta Distribution of Proportion P

I show the average of function of random variable's quantile values (1) and moments of the beta random variable (2), before deriving mean, variance and the approximate beta distribution of proportion P (3).

(1) The Asymptotic Average of Function of Random Variable's Quantile Values

Let $Q(Y|q)$ the q quantile value of random variable Y, whose probability density function is $p(Y)$. Then,

$$\begin{aligned}
\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} f\left[Q\left(Y\middle|\frac{i}{n}\right)\right] &= \lim_{n\to\infty}\sum_{i=1}^{n} f\left[Q\left(Y\middle|\frac{i}{n}\right)\right]\frac{\frac{1}{n}}{Q\left(Y\middle|\frac{i}{n}\right)-Q\left(Y\middle|\frac{i-1}{n}\right)} \\
&\quad \left[Q\left(Y\middle|\frac{i}{n}\right)-Q\left(Y\middle|\frac{i-1}{n}\right)\right] \\
&= \lim_{n\to\infty}\sum_{i=1}^{n} f[Q(Y|q_i)]\frac{dq[Q(Y|q_i)]}{dQ(Y|q_i)}dQ(Y|q_i) \\
&= \int_{Y_{min}}^{Y_{max}} f(Y)p(Y)dY \\
&= E[f(Y)] \qquad (1)
\end{aligned}$$

where

$$\begin{aligned}
q_i &= \frac{i}{n} \\
dq[Q(Y|q_i)] &= q_i - q_{i-1} = \frac{1}{n} \\
dQ(Y|q_i) &= Q\left(Y\middle|\frac{i}{n}\right)-Q\left(Y\middle|\frac{i-1}{n}\right)
\end{aligned}$$

Thus, the average of function of random variable's quantile values is asymptotically equal to expectation of the function.

(2) Moments of the Beta Random Variable

Suppose that $\pi \sim \mathcal{BT}(\pi|\bar{\pi}, \gamma)$. It follows that

$$\begin{aligned}
E(\pi^d) &= \int_0^1 \pi^d \frac{\Gamma(\gamma^{-1})}{\Gamma(\bar{\pi}\gamma^{-1})\Gamma([1-\bar{\pi}]\gamma^{-1})} \pi^{(\bar{\pi}\gamma^{-1})-1}(1-\pi)^{([1-\bar{\pi}]\gamma^{-1})-1} d\pi \\
&= \frac{\Gamma(\gamma^{-1})}{\Gamma(\bar{\pi}\gamma^{-1})\Gamma([1-\bar{\pi}]\gamma^{-1})} \frac{\Gamma(\bar{\pi}\gamma^{-1}+d)\Gamma([1-\bar{\pi}]\gamma^{-1})}{\Gamma(\gamma^{-1}+d)} \\
&\quad \int_0^1 \frac{\Gamma(\gamma^{-1}+d)}{\Gamma(\bar{\pi}\gamma^{-1}+d)\Gamma([1-\bar{\pi}]\gamma^{-1})} \pi^{(\bar{\pi}\gamma^{-1})+d-1}(1-\pi)^{([1-\bar{\pi}]\gamma^{-1})-1} d\pi \\
&= \frac{\Gamma(\gamma^{-1})}{\Gamma(\bar{\pi}\gamma^{-1})} \frac{\Gamma(\bar{\pi}\gamma^{-1}+d)}{\Gamma(\gamma^{-1}+d)} \\
&\quad \int_0^1 \mathcal{BT}\left[\pi \middle| \frac{\bar{\pi}\gamma^{-1}+d}{\gamma^{-1}+d}, \left(\frac{1}{\gamma^{-1}+d}\right)\right] d\pi \\
&= \frac{\Gamma(\bar{\pi}\gamma^{-1}+d)}{\Gamma(\bar{\pi}\gamma^{-1})} \left(\frac{\Gamma(\gamma^{-1}+d)}{\Gamma(\gamma^{-1})}\right)^{-1}
\end{aligned}$$

$$\therefore E(\pi^1) = \bar{\pi} \qquad \text{(Eq. (2))}$$

$$E(\pi^2) = \frac{\bar{\pi}\gamma^{-1}+1}{\gamma^{-1}+1}\bar{\pi} \qquad \text{(Eq. (3))}$$

$$\begin{aligned}
V(\pi) &= E(\pi^2) - [E(\pi)]^2 \\
&= \frac{\bar{\pi}(1-\bar{\pi})}{1+\gamma^{-1}} \qquad \text{(Eq. (4))}
\end{aligned}$$

(3) Mean, Variance and the Approximate Beta Distribution of Proportion P

Without specifying distribution of P, its mean and variance are;

$$
\begin{aligned}
E(P) &= E\left(\frac{1}{n}\sum_{i=1}^{n} Z_i\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} E(Z_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} \pi_i \\
\therefore \lim_{n\to\infty} E(P) &\to \int_0^1 \pi \mathcal{BT}(\pi|\bar{\pi},\gamma)d\pi \qquad (\because \text{Eq. (1)}) \\
&= \bar{\pi} \qquad (\text{Eq. (5)}) \\
V(P) &= V\left(\frac{1}{n}\sum_{i=1}^{n} Z_i\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} V(Z_i) \qquad (\because \text{Cor}(Z_i, Z_j) = 0 \quad \text{for} \quad i \neq j) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} (\pi_i - \pi_i^2) \\
\therefore \lim_{n\to\infty} V(P) &\to \frac{n}{n^2}\int_0^1 (\pi - \pi^2)\mathcal{BT}(\pi|\bar{\pi},\gamma)d\pi \qquad (\because \text{Eq. (1)}) \\
&= \frac{\bar{\pi}(1-\bar{\pi})}{n(1+\gamma)} \qquad (\because \text{Eq. (2) and (3)}) \qquad (\text{Eq. (6)})
\end{aligned}
$$

Now approximate the distribution of P by beta density function;

$$P \sim \mathcal{BT}(p|\bar{p}, g)$$

For sufficiently large n, we can regard;

$$\begin{aligned}
\bar{p} &\equiv E(P) (\because \text{Eq. (2)}) \\
&= \bar{\pi} \qquad (\because \text{Eq. (5)}) \\
V(P) &= \frac{\bar{p}(1-\bar{p})}{1+g^{-1}} \qquad (\because \text{Eq. (4)}) \\
V(P) &= \frac{\bar{\pi}(1-\bar{\pi})}{n(1+\gamma)} \qquad (\because \text{Eq. (6)}) \\
\therefore g &= \frac{1}{n(1+\gamma)-1}
\end{aligned}$$

References

Binder, Sarah A. 2003. *Stalemate: Causes and Consequences of Legislative Gridlock*. Washington D. C.: Brookings Institution Press.

Palmquist, Bradey. 1997. Heterogeneity and Dispersion in the Beta-Binomial Model. In *the Annual Meeting of the American Political Science Association*. Washington D.C.: .

Paolino, Philip. 2001. "Maximum Likelihood Estimation of Models with Beta-Distributed Dependent Variable." *Political Analysis* 9:325–46.